

**IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
BEFORE THE BOARD OF PATENT APPEALS AND INTERFERENCES**

Application of:	Daniel Dulitz, et al.	Confirmation No.:	7663
Serial No.:	10/614,111	Art Unit:	2168
Filed:	July 3, 2003	Examiner:	Morrison, Jay A.
For:	<i>Duplicate Document Detection in a Web Crawler System</i>	Attorney Docket No.:	60963-0005-US

Mail Stop Appeal Brief Patents
Commissioner for Patents
Arlington, VA 22202

APPELLANTS' BRIEF UNDER 37 C.F.R. § 41.37

This Appeal Brief in conjunction with the previously filed Notice of Appeal appeals the §103 rejection of claims 12-20, 37-40, and 42-58 by the United States Patent and Trademark Office in a fourth Office Action dated August 20, 2007.

Appellants demonstrate that independent claims 12, 18, 37, 40, 50, and 56 have a claim limitation not taught by the cited references Meyerzon, Cho, or Rujan, and thus the rejection under 35 U.S.C. § 103 cannot be sustained.

The fee required under 37 C.F.R. § 1.17(c) is being filed concurrently herewith.

Table of Contents

I.	THE REAL PARTY IN INTEREST	3
II.	RELATED APPEALS AND INTERFERENCES.....	4
III.	STATUS OF THE CLAIMS	5
IV.	STATUS OF AMENDMENTS.....	6
V.	SUMMARY OF THE CLAIMED SUBJECT MATTER	7
	A. The Subject Matter as Claimed in Independent Claim 12	7
	B. The Subject Matter of Independent Claims 18, 37, 40, 50, and 56	9
VI.	GROUND OF REJECTION PRESENTED FOR REVIEW	11
	A. The § 103 Rejection of Claims 12-17, 40, 42-48, and 50-55.....	11
	B. The § 103 Rejection of Claims 18–20, 37–39, and 56–58.....	12
	C. The § 103 Rejection of Claim 49.....	13
	D. Summary of Rejections.....	14
VII.	ARGUMENT	15
	A. To reject claims under 35 U.S.C. § 103, all claim limitations must be taught.	15
	B. A “representative document” is the one indexed and thus presented to a user....	15
	C. The claims require that the representative document changes for some documents.	17
	D. Meyerzon does not teach a web crawling methodology where the representative document can change.	17
	E. Cho does not teach a web crawling methodology where the representative document can change.....	18
	F. Rujan does not teach a web crawling methodology where the representative document can change.....	19
	G. Lambert does not teach a web crawling methodology where the representative document can change.....	19
	H. Conclusion	19
VIII.	Claims Appendix	21
IX.	Evidence Appendix	30
X.	Related Proceedings Appendix	31

I. THE REAL PARTY IN INTEREST

The real party in interest in this appeal is Google Inc., the assignee of this application.

II. RELATED APPEALS AND INTERFERENCES

Appellants are not aware of any appeals, judicial proceedings, or interferences that will affect directly, will be affected directly by, or will otherwise have a bearing on, the decision in this appeal.

III. STATUS OF THE CLAIMS

The status of the claims is as follows:

- Claims canceled: 1–11, 21–36, 41.
- Claims withdrawn from consideration but not cancelled: None.
- Claims pending: 12–20, 37–40, 42–58.
- Claims rejected: 12–20, 37–40, 42–58.
- Claims appealed: 12–20, 37–40, 42–58.

IV. STATUS OF AMENDMENTS

All amendments have been entered. A copy of the appealed claims is attached as Section VIII, "Claims Appendix."

V. SUMMARY OF THE CLAIMED SUBJECT MATTER

This application has six pending independent claims, and each of these independent claims incorporates the subject matter described.

A. The Subject Matter as Claimed in Independent Claim 12

The claimed subject matter of claim 12 is a method of handling duplicate documents in a network crawling system.¹ In the claimed method, when there are duplicates a representative document is selected for the set of duplicates, and over time the representative may change.²

Initially, a set of tables is created that stores information about documents on a network.³ This information includes what documents are duplicates, and the rank (or score) of each document.⁴ The rank generally indicates the importance or popularity of each document.⁵

“Crawling” new documents comprises several operations.⁶ Briefly, the method entails comparing the new document to an existing set of documents having the same content.⁷ The new document becomes the representative for that content under certain circumstances.⁸ If the new document becomes the representative, it is indexed.⁹

The claims and specification describe the crawling operation in more detail. A new document is received, and has two properties: its “document identifier” and a “document

¹ Application specification ¶ [0006].

² Application specification ¶¶ [0006], [0008].

³ Application specification ¶¶ [0024], [0026], [0051], [0052]; elements 324, 326, 328, 340, 342, 344 in Fig. 3, elements 340, 342, and 344 in Fig. 4.

⁴ Application specification ¶¶ [0051], [0052]; *see, e.g.*, elements 3410-1 to 3410-4 in Fig. 4.

⁵ Application specification ¶¶ [0007], [0044].

⁶ Application specification ¶¶ [0006] – [0008]; Figs. 5–10 (providing flowcharts).

⁷ Application specification ¶¶ [0006] – [0008]; elements 1470-20, 1470-30, 1470-50, and 1470-60 in Fig. 7.

⁸ Application specification ¶¶ [0006] – [0008], [0048]; elements 1470-30, 1470-50, 1470-60, 1470-70 of Fig. 7. The representative is sometimes referred to as the “canonical page.”

⁹ Application specification ¶¶ [0069], [0070]. *See also* claim 12 and specification paragraphs [0002] and [0023], indicating that indexing makes the page available to a search engine.

rank.”¹⁰ The document identifier identifies the content of the document.¹¹ For example, the specification teaches using a 64 bit “content fingerprint.”¹² The document rank is a numeric ranking of the document compared to other documents on the network.¹³

Another operation is to identify other known documents that share the same content as the new document.¹⁴ This data is read from the tables.¹⁵ This set of documents has a representative document, which the claims refer to as the “original representative document.”¹⁶

Once the new document and the set of documents are identified, the method calls for determining a representative of the enlarged set.¹⁷ The enlarged set comprises the set of documents identified earlier, together with the new document.¹⁸ The representative of the enlarged set may be different from the original representative document because the representative of the enlarged set may be the new document.¹⁹ The specification provides examples of determining the representative based on the rankings of the documents.²⁰ One method also applies a hysteresis test so that a change in the representative occurs only when the new document is sufficiently better than the original representative document.²¹

The information in the tables is updated to include the new document, and potentially changes to other documents.²² For example, if the document that was the original representative is no longer the representative, the table is updated to show that fact.²³

¹⁰ Application specification ¶¶ [0047], [0062]; element 1450-2 in Fig. 5 and element 1470-10 in Fig. 7.

¹¹ Application specification ¶¶ [0007], [0047].

¹² Application specification ¶¶ [0007], [0047].

¹³ Application specification ¶¶ [0007], [0044].

¹⁴ Application specification ¶ [0007], [0067], [0068]; element 1470-20 in Fig. 7

¹⁵ Application specification ¶¶ [0007], [0051] (data stored in a content fingerprint table (CFT)), [0066] – [0068]; element 340 in Fig. 4.

¹⁶ Application specification ¶¶ [0067], [0068]; elements 3410-3 and 3420-2 in Fig. 4. See also claim 12.

¹⁷ Application specification ¶¶ [0006], [0008]; elements 1470-30, 1470-60, 1470-65, 1470-70, and 1470-80 in Fig. 7.

¹⁸ Application specification ¶¶ [0006], [0008].

¹⁹ Application specification ¶¶ [0006], [0008]; element 1470-70 in Fig. 7.

²⁰ Application specification ¶¶ [0008], [0067].

²¹ Application specification ¶¶ [0068], [0069]; element 1470-60 in Fig. 7.

²² Application specification ¶¶ [0069] – [0074]; elements 1470-60, 1470-65, and 1470-70 in

In addition, if the new document becomes the representative document, it is indexed.²⁴ Indexing is the operation that makes the document available to a search engine.²⁵

The crawling operation described above is repeated for multiple documents, and in some cases the new document becomes the representative document.²⁶ When the new document becomes the representative document, the representative document has changed because the original representative document was from a set of documents that did not include the new document.²⁷

B. The Subject Matter of Independent Claims 18, 37, 40, 50, and 56

Much of the subject matter taught in independent claim 12 also appears in the other independent claims. Because the subject matter of claim 12 is sufficient to establish patentability, Appellants believe that it is unnecessary to substantially repeat the above description. Therefore, only the aspects of the other independent claims that have different support in the specification than claim 12 are described here.

Claim 18 additionally requires $N+1$ tables, where N is an integer greater than one, wherein the $N+1$ tables comprise N tables, each generated during a respective phase of a set of N crawling phases, and a current table generated during a current one of the N crawling phases, wherein an oldest one of the N tables was generated during a previous instance of the current crawling phase.²⁸ The method of claim 18 includes reading information stored in the $N+1$ tables to identify a set of documents sharing the document identifier of the newly crawled document.²⁹ Upon completion of a current crawling phase, the oldest one of the N tables is retired.³⁰

System claim 37 has the same additional requirements described above for claim 18, as well as one or more central processing units and a network interface.³¹

Fig. 7, Fig. 9 (flowchart showing the table update operation).

²³ Application specification ¶ [0070] (unmarking entries as necessary to show that they are no longer the canonical pages).

²⁴ Application specification ¶¶ [0069], [0070]. *See also* claim 12.

²⁵ Application specification ¶¶ [0002], [0023].

²⁶ See claim 12, last paragraph.

²⁷ Application specification ¶ [0070]; element 1470-70 in Fig. 7.

²⁸ Application specification ¶¶ [0078]

²⁹ Application specification ¶ [0079].

³⁰ Application specification ¶ [0078].

³¹ Application specification ¶¶ [0051]; elements 302 and 310 in Fig. 3.

Claim 40, directed to a computer program product, has elements that have the same support in the specification as for claim 12.

Claim 50, directed to a computer program product, has elements that have the same support in the specification as for claim 12.

Claim 56, directed to a computer program product, has elements that have the same support in the specification as for claim 18.

VI. GROUNDS OF REJECTION PRESENTED FOR REVIEW

Examiner Morrison has rejected all of the pending claims under 35 U.S.C. § 103.

A. The § 103 Rejection of Claims 12-17, 40, 42-48, and 50-55.

The Examiner stated in the Office Action mailed 08/20/2007 on page 2 that:

Claims 12-17, 40, 42-48 and 50-55 are rejected under 35 U.S.C. 103(a) as being unpatentable over Meyerzon et al. ('Meyerzon' hereinafter) (Patent Number 6,547,829) in view of Cho et al. ('Cho' hereinafter) ("Finding replicated web collections," by Cho et al., Proceedings of the ACM SIGMOD International Conference on Management of Data, pages 355–366, 2000).

As per claim 12, Meyerzon teaches

A method of detecting duplicate documents in a network crawling system, comprising: (see abstract and background)

constructing a plurality of tables, each table corresponding to a portion of a document address space (builds new index based on documents, column 4, lines 43–60), storing information identifying documents having a same document identifier and each identified document having an associated document rank; (column 2, lines 3-16)

receiving a newly crawled document, such document characterized by a document identifier and a document rank; (column 2, lines 3-16)

reading information stored in the plurality of tables to identify a set of documents, sharing the document identifier of the newly crawled document, and ascertaining as original representative document for the identified set of documents; (column 9, lines 18-29)

updating the information stored in at least one of the tables in accordance with the document ranks of the identified set of documents and the newly crawled document; (column 2, lines 3-16)

determining a representative document for the newly crawled documents and the identified set of documents (column 9, lines 32-40)

Meyerzon does not explicitly indicate "indexing the representative document when the representative document is the newly crawled document; and repeating the receiving, reading, updating, determining and indexing operations with respect to a plurality of newly crawled documents, each of which shares a respective document identifier with a respective set of

documents, such that at least some of the newly crawled documents are determined to be representative documents and are indexed”.

However, Cho discloses “indexing the representative document when the representative document is the newly crawled document; and repeating the receiving, reading, updating, determining and indexing operations with respect to a plurality of newly crawled documents, each of which shares a respective document identifier with a respective set of documents, such that at least some of the newly crawled documents are determined to be representative documents and are indexed” (newly replicated collection, page 365, first column, second paragraph; one page displayed or represents collection of duplicate document, page 365, second column, first paragraph).

It would have been obvious to one of skill in the art at the time the invention was made to combine Meyerzon and Cho because using the steps of “indexing the representative document when the representative document is the newly crawled document; and repeating the receiving, reading, updating, determining and indexing operations with respect to a plurality of newly crawled documents, each of which shares a respective document identifier with a respective set of documents, such that at least some of the newly crawled documents are determined to be representative documents and are indexed” would have given those skilled in the art the tools to improve the invention by allowing duplicate documents to be identified and represented. This gives the use the advantage of not having multiple copies of the same document to choose from .

For this appeal, Appellants focus on the limitation “repeating the receiving, reading, updating, determining and indexing operations with respect to a plurality of newly crawled documents, each of which shares a respective document identifier with a respective set of documents, such that at least some of the newly crawled documents are determined to be representative documents and are indexed.” This limitation is a portion of what the Examiner indicated was not taught by Meyerzon. Appellants will demonstrate that this limitation is not taught by Meyerzon or Cho.

Appellants seek to streamline the appeal process by focusing on this limitation, but do not thereby admit to the correctness or appropriateness of any other statement or issue raised by Examiner Morrison.

B. The § 103 Rejection of Claims 18–20, 37–39, and 56–58.

The Examiner stated in the Office Action mailed 08/20/2007 on page 11 that:

Claims 18-20, 37-39 and 56-58 are rejected under 35 U.S.C. 103(a) as being unpatentable over Meyerzon et al. ('Meyerzon' hereinafter) (Patent Number 6,547,829) in view of Cho et al. ('Cho' hereinafter) ("Finding replicated web collections," by Cho et al., Proceedings of the ACM SIGMOD International Conference on Management of Data, pages 355-366, 2000) and further in view of Rujan et al. ('Rujan' hereinafter) (Patent Number 6,976,207)

The Examiner's analysis of independent claim 18 is similar to that in claim 12 above, and includes the same statements regarding the limitation "indexing the representative document when the representative document is the newly crawled document; and repeating the receiving, reading, updating, determining and indexing operations with respect to a plurality of newly crawled documents, each of which shares a respective document identifier with a respective set of documents, such that at least some of the newly crawled documents are determined to be representative documents and are indexed."

As above, Appellants focus on the limitation "repeating the receiving, reading, updating, determining and indexing operations with respect to a plurality of newly crawled documents, each of which shares a respective document identifier with a respective set of documents, such that at least some of the newly crawled documents are determined to be representative documents and are indexed," making no admission regarding the correctness or appropriateness of any other statement or issue raised by Examiner Morrison. Appellants will demonstrate that this limitation is not taught by Meyerzon, Cho, or Rujan.

C. The § 103 Rejection of Claim 49.

The Examiner cited a fourth reference against dependent claim 49. Specifically, the Examiner, in the Office Action mailed 08/20/2007 at page 15, stated that:

Claim 49 is rejected under 35 U.S.C. 103(a) as being unpatentable over Meyerzon et al. ('Meyerzon' hereinafter) (Patent Number 6,547,829) in view of Cho et al. ('Cho' hereinafter) ("Finding replicated web collections," by Cho et al., Proceedings of the ACM SIGMOD International Conference on Management of Data, pages 355-366, 2000) and further in view of Lambert et al. ('Lambert' hereinafter) (Patent Number 6,976,207 [sic, Patent Application Publication 2002/0038350]).

The Examiner cited Lambert to address a specific limitation in dependent claim 49, which Appellants are not addressing at this time. As above, Appellants focus on the limitation "repeating the receiving, reading, updating, determining and indexing operations

with respect to a plurality of newly crawled documents, each of which shares a respective document identifier with a respective set of documents, such that at least some of the newly crawled documents are determined to be representative documents and are indexed,” making no admission regarding the correctness or appropriateness of any other statement or issue raised by Examiner Morrison. Appellants will demonstrate that this limitation is not taught by Meyerzon, Cho, or Lambert.

D. Summary of Rejections

The limitation “repeating the receiving, reading, updating, determining and indexing operations with respect to a plurality of newly crawled documents, each of which shares a respective document identifier with a respective set of documents, such that at least some of the newly crawled documents are determined to be representative documents and are indexed” appears in each of the independent claims, and has been rejected on the same basis. In each case the Examiner asserted that this limitation is taught by Cho.

VII. ARGUMENT

Appellants argue that the limitation “repeating the receiving, reading, updating, determining and indexing operations with respect to a plurality of newly crawled documents, each of which shares a respective document identifier with a respective set of documents, such that at least some of the newly crawled documents are determined to be representative documents and are indexed” is not taught by any of the asserted references Meyerzon, Cho, Rujan, or Lambert.

A. To reject claims under 35 U.S.C. § 103, all claim limitations must be taught.

Case law requires that to “establish *prima facie* obviousness of a claimed invention, all the claim limitations must be taught or suggested by the prior art.” *In re Royka*, 490 F.2d 981, 180 USPQ 580 (CCPA 1974) as cited at MPEP 2143.03.

B. A “representative document” is the one indexed and thus presented to a user.

Indexing is the operation that makes documents available to a search engine.³² By indexing a representative document when there are duplicates, the system saves processing resources.³³ In addition, indexing a representative document from each set of duplicates provides a better user experience in response to a query: diverse results are not crowded out by duplicates.³⁴ Indexing a representative document from each set of duplicates is how the invention achieves its results.

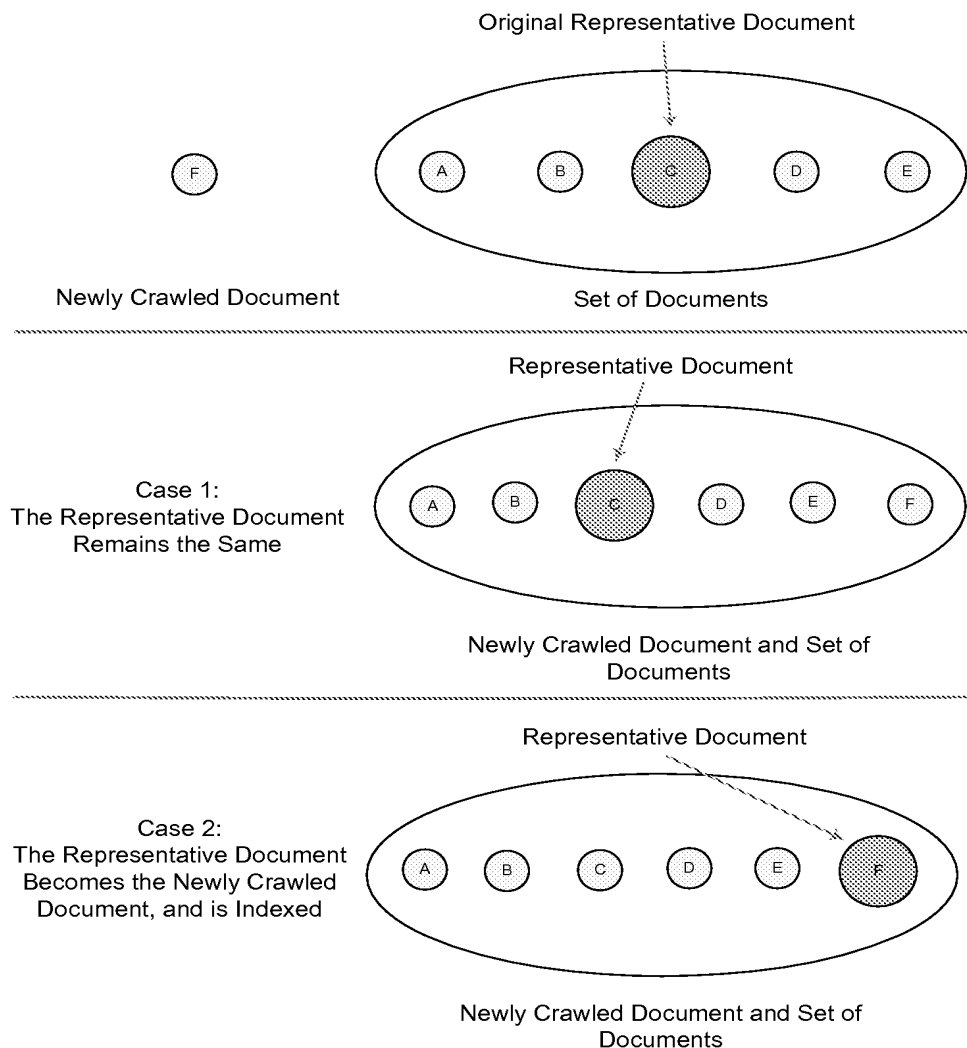
Indexing a representative document is recited in the claims. The claims require “determining a representative document for the newly crawled document and the identified set of documents” and “indexing the representative document when the representative document is the newly crawled document.” This makes sense. The documents in the “identified set of documents” all have the same document content as the newly crawled document. If the newly crawled document becomes the representative, then it needs to be

³² Application specification ¶¶ [0002], [0023].

³³ Application specification ¶ [0004].

indexed; but if the representative document stays the same, the newly crawled document does not need to be indexed.

The diagram below depicts this process graphically. The top portion shows newly crawled document F, and the set of documents A, B, C, D, and E, all sharing the same content as document F. Here, document C is shown as the original representative document for documents A, B, C, D, and E. Next, document F is added to the set, and a representative is selected. The middle portion of the diagram shows the case where document C remains the representative. The bottom portion of the diagram shows the case where document F becomes the representative. In this case document F is indexed.



³⁴ Application specification ¶ [0004].

C. The claims require that the representative document changes for some documents.

The claim language “such that at least some of the newly crawled documents are determined to be representative documents and are indexed” conveys the point that the representative document changes for some of the documents. Importantly, the claims address the case where the newly crawled documents are duplicates of documents already known, so selecting the newly crawled document as the representative changes the representative.

For each newly crawled document, the reading operation identifies an original representative document with the same content³⁵ as the newly crawled document. The original representative document is not the newly crawled document because the original representative document was ascertained from a set of documents already stored in tables.

Therefore, when newly crawled documents are “determined to be representative documents and are indexed,” the representative documents have changed.

D. Meyerzon does not teach a web crawling methodology where the representative document can change.

Meyerzon addresses the detection of duplicate documents, but responds with a “first copy wins” approach. Meyerzon explains this in the specification at column 9, lines 33-40, with reference to figure 3. When a document is crawled, the crawler determines if the content of the newly crawled document matches the content of a document already in the history table.³⁶ If the content already exists, then the address (URL) of the newly crawled document is just saved to the history table. If it is not found, then several steps are performed, including step S25, which indexes the new document. Because the first copy of a document is always the one that is indexed, there is no discussion of representative documents, or changing the representative document.

In addition, the Examiner pointed out that Meyerzon does not teach the limitation addressed in this appeal. In the Office Action mailed 08/20/2007, the Examiner stated on

³⁵ The claim language literally says “sharing the document identifier.” Paragraph [0004] in the specification explains: “In one embodiment, a document identifier is a fixed length fingerprint of a document’s content.”

³⁶ Meyerzon uses a “CID,” which is defined as a “content identifier.” See abstract; column 2, lines 65-67.

page 4 that “Meyerzon does not explicitly indicate “indexing the representative document when the representative document is the newly crawled document; and repeating the receiving, reading, updating, determining and indexing operations with respect to a plurality of newly crawled documents, each of which shares a respective document identifier with a respective set of documents, such that at least some of the newly crawled documents are determined to be representative documents and are indexed”.

E. Cho does not teach a web crawling methodology where the representative document can change.

Cho teaches detection of duplicate documents, but the first copy found remains the permanent representative. Cho refers to this methodology as a “replica avoiding process.”³⁷ Specifically, “each crawl identifies new replicated collections that can be avoided in the future.”³⁸ Like Meyerzon, the first copy discovered is the one that is indexed and used. This section of Cho does not teach changing and indexing a representative document.

Cho also presents a revised way to display query results, but does not suggest determining and indexing a new representative.³⁹ Cho teaches that it is useful to continue gathering multiple copies of document collections because one of the copies may be unavailable later.⁴⁰ In response to a user query, Cho discloses a “presentation filter” that “rolls up” collections so that “it only displays the link of one page in a collection, even if multiple pages within the collection satisfy the query.”⁴¹ Thus, Cho keeps a record of duplicate documents which can be presented to a user, but only one is indexed and in the normal course only one is presented to the user.

The Examiner refers to §§ 5.1, 5.2 of Cho,⁴² which do not teach changing and indexing the representative for a set of duplicate documents. The Examiner first refers to “newly replicated collection, page 365, first column, second paragraph.” This is § 5.1 of Cho, which discloses only a replica avoiding process. The first copy is indexed, and remains the representative permanently. The Examiner then refers to “one page displayed or

³⁷ Cho, last paragraph of § 5.1.

³⁸ Cho, last paragraph of § 5.1.

³⁹ Cho § 5.2, ¶ 1.

⁴⁰ Cho § 5.2, ¶ 1.

⁴¹ Cho § 5.2, ¶ 5.

⁴² Office Action mailed 08/20/2007 at page 4.

represents collection of duplicate document, page 365, second column, first paragraph.” This is § 5.2 of Cho, which teaches only a presentation filter, as described above. Thus, neither § 5.1 nor § 5.2 teach having a set of duplicate documents where there is a representative document that changes and is indexed.

F. Rujan does not teach a web crawling methodology where the representative document can change.

The Examiner argued that Rujan teaches a limitation regarding “retiring the oldest one,” which is not the limitation addressed in this appeal.⁴³ Rujan teaches a classification method, which is not relevant to the claim limitation addressed here. Most importantly, Rujan contains no discussion of duplicate documents. Without a discussion of duplicate documents, there is no discussion of representative documents or changing representative documents.

G. Lambert does not teach a web crawling methodology where the representative document can change.

The Examiner argued that Lambert teaches a limitation regarding “a document is a temporary redirect page ...,” which is not the limitation addressed in this appeal.⁴⁴ Lambert teaches a method of “enhanced web page delivery,” which Lambert says “employs techniques in identifying visitors, both humans and search engine spiders, and appropriately redirecting them to specific Universal Resource Locators.” The subject matter in Lambert is not relevant to the claim limitation addressed here. Most importantly, Lambert contains no discussion of duplicate documents. Without a discussion of duplicate documents, there is no discussion of representative documents or changing representative documents.

H. Conclusion

In summary, Appellants have demonstrated that the § 103 rejections cannot be sustained because the combination of references does not teach the claim limitation “repeating the receiving, reading, updating, determining and indexing operations with respect to a plurality of newly crawled documents, each of which shares a respective document

⁴³ Appellants do not admit that Rujan teaches the limitation as stated by the Examiner, but do not address that issue in this appeal.

⁴⁴ Appellants do not admit that Lambert teaches the limitation as stated by the Examiner, but do not address that issue in this appeal.

identifier with a respective set of documents, such that at least some of the newly crawled documents are determined to be representative documents and are indexed,” which appears in all of the pending claims.

In view of the foregoing, Appellants respectfully request the reversal of Examiner Morrison’s rejections. Appellants further request allowance of the pending claims 12-20, 37-40, 42-58. If there are any other fees due in connection with the filing of this Brief, please charge the fees to Morgan, Lewis & Bockius LLP Deposit Account No. 50-0310 (order no. 60963-0005-US).

If a fee is required for an extension of time under 37 C.F.R. § 1.136 not accounted for above, such an extension is requested and the fee should be charged to Morgan, Lewis & Bockius LLP Deposit Account No. 50-0310 (order no. 60963-0005-US).

Respectfully submitted,

MORGAN, LEWIS & BOCKIUS LLP

Dated: March 20, 2008

By: / Gary S. Williams /
Gary S. Williams
Reg. No. 31,066

Customer No.: 24341
MORGAN, LEWIS & BOCKIUS LLP
3000 El Camino Real, Suite 700
Palo Alto, CA 94306
650-843-4000

VIII. Claims Appendix

CLAIMS CURRENTLY ON APPEAL ORDERED BY NUMBER

1 – 11. (Canceled).

12. A method of detecting duplicate documents in a network crawling system, comprising:

constructing a plurality of tables, each table corresponding to a portion of a document address space, storing information identifying documents having a same document identifier and each identified document having an associated document rank;

receiving a newly crawled document, such document characterized by a document identifier and a document rank;

reading information stored in the plurality of tables to identify a set of documents sharing the document identifier of the newly crawled document, and ascertaining an original representative document for the identified set of documents;

updating the information stored in at least one of the tables in accordance with the document ranks of the identified set of documents and the newly crawled document;

determining a representative document for the newly crawled document and the identified set of documents;

indexing the representative document when the representative document is the newly crawled document; and

repeating the receiving, reading, updating, determining and indexing operations with respect to a plurality of newly crawled documents, each of which shares a respective document identifier with a respective set of documents, such that at least some of the newly crawled documents are determined to be representative documents and are indexed.

13. The method of claim 12, wherein information identifying the identified set of documents, including a particular document serving as the original representative document of the identified set, is stored in one or more tables.

14. The method of claim 13, wherein the determining includes

comparing the document rank of the newly crawled document with that of the particular document from the identified set in accordance with a set of predefined comparison criteria;

selecting the newly crawled document as the representative document if the set of predefined comparison criteria are met; and

keeping the particular document as the representative document if the set of predefined comparison criteria is not met.

15. The method of claim 14, wherein the set of predefined comparison criteria comprise at least two parameters, one parameter for comparison with an absolute difference of document ranks between the newly crawled document and the particular document, and another parameter for comparison with a ratio of document ranks between the newly crawled document and the particular document.

16. The method of claim 12, wherein the updating includes inserting information identifying the newly crawled document into the at least one table only when a predefined insertion condition is satisfied.

17. The method of claim 16, wherein the predefined insertion condition is that the document rank of the newly crawled document is higher than the document rank of at least one document in the identified set of documents.

18. A method of detecting duplicate documents in a network crawling system, comprising:

constructing a plurality of tables, each table corresponding to a segment of a document address space, storing information identifying documents having a same document identifier and each identified document having an associated document rank, wherein the plurality of tables comprise $N+1$ tables where N is an integer greater than one, wherein the $N+1$ tables comprise N tables, each generated during a respective phase of a set of N crawling phases, and a current table generated during a current one of the N crawling phases, wherein an oldest one of the N tables was generated during a previous instance of the current crawling phase;

receiving a newly crawled document, such document characterized by a document identifier and a document rank;

reading information stored in the N+1 tables to identify a set of documents sharing the document identifier of the newly crawled document, and ascertaining an original representative document for the identified set of documents;

updating the information stored in the current table in accordance with the document rankings of the identified set of documents and the newly crawled document;

determining a representative document for the newly crawled document and the identified set of documents;

indexing the representative document when said representative document is the newly crawled document;

repeating the receiving, reading, updating, determining and indexing operations with respect to a plurality of newly crawled documents, each of which shares a respective document identifier with a respective set of documents, such that at least some of the newly crawled documents are determined to be representative documents and are indexed; and

upon completion of the current crawling phase, retiring the oldest one of the N tables.

19. The method of claim 18, wherein the reading comprises reading from a merged table that stores information from a plurality of the N tables, and reading from the current table.

20. The method of claim 18, wherein information identifying the identified set of documents, including a particular document serving as the original representative document of the identified set, is stored in one or more tables.

21 – 36. (Canceled).

37. A system for detecting duplicate documents during network crawling, comprising:
one or more central processing units for executing programs;
a network interface for receiving documents; and
a duplicate document detection engine executable by the one or more central processing units, the engine comprising:

a plurality of tables, each table corresponding to a segment of a document address space, storing information identifying documents having a same document identifier and each identified document having an associated document rank, wherein the plurality of tables comprise N+1 tables where N is an integer greater than one, wherein the N+1 tables comprise N tables, each generated during a respective phase of a set of N crawling phases, and a current table generated during a current one of the N crawling phases, wherein an

oldest one of the N tables was generated during a previous instance of the current crawling phase;

instructions for receiving a newly crawled document, such document characterized by a document identifier and a document rank;

instructions for reading information stored in the N+1 tables to identify a set of documents, sharing the document identifier of the newly crawled document, and ascertaining an original representative document for the identified set of documents;

instructions for updating the information stored in the current table in accordance with the document rankings of the identified set of documents and the newly crawled document;

instructions for determining a representative document for the newly crawled document and the identified set of documents;

instructions for indexing the representative document when said representative document is the newly crawled document;

instructions for repeating the receiving, reading, updating, determining and indexing operations with respect to a plurality of newly crawled documents, each of which shares a respective document identifier with a respective set of documents, such that at least some of the newly crawled documents are determined to be representative documents and are indexed; and

instructions for retiring the oldest one of the N tables upon completion of the current crawling phase.

38. The system of claim 37 wherein the reading comprises reading from a merged table that stores information from a plurality of the N tables, and reading from the current table.

39. The system of claim 37, wherein the identified set of documents, including a particular document serving as the original representative document of the identified set, are stored in one or more tables.

40. A computer program product for use in conjunction with a computer system, the computer program product comprising a computer readable storage medium and a computer program mechanism embedded therein, the computer program mechanism comprising:

instructions for constructing a plurality of data structures for storing information of documents, each document characterized by a document identifier and a document rank, the

information stored in the plurality of data structures include the document identifier and a document rank for each document;

instructions for receiving a requesting document in association with its document identifier and document rank;

instructions for selecting from the plurality of data structures a set of documents sharing the same document identifier as the requesting document, and ascertaining an original representative document for the identified set of documents;

instructions for generating a new set of documents from the requesting document and the selected set of documents in accordance with their document rank;

instructions for identifying a representative document of the new set of documents;

instructions for indexing the representative document when said representative document is the newly crawled document; and

instructions for repeating the receiving, reading, updating, determining and indexing operations with respect to a plurality of newly crawled documents, each of which shares a respective document identifier with a respective set of documents, such that at least some of the newly crawled documents are determined to be representative documents and are indexed.

41. (Canceled).

42. The computer program product of claim 40, wherein the plurality of data structures include a data structure for storing information of multiple sets of documents, each set of documents sharing a same document content.

43. The computer program product of claim 40, wherein the plurality of data structures include a data structure for storing information of multiple sets of documents, each set of documents sharing a same document address.

44. The computer program product of claim 40, wherein the document identifier is a fixed length fingerprint of document content of a document characterized by the document identifier.

45. The computer program product of claim 40, wherein the document identifier is a fixed length fingerprint of an address of a document characterized by the document identifier.

46. The computer program product of claim 40, wherein the generating instructions include

sorting the requesting document and the selected set of documents in accordance with a metric included in score information of the requesting document and selected set of documents; and

selecting a new set of documents, having at most a predefined number of documents, from the requesting document and the selected set of documents based on the sorting result.

47. The computer program product of claim 40, wherein the score information for each document includes a document rank; and the identifying instructions include

comparing the document rank of the requesting document with that of a particular document from the selected set of documents in accordance with a set of predefined comparison criteria, wherein the particular document was previously determined to be the representative document for the selected set of documents;

selecting the requesting document as the representative document for the new set of documents if the set of predefined comparison criteria are met; and

keeping the particular document as the representative document for the new set of documents if the set of predefined comparison criteria is not met.

48. The computer program product of claim 47, wherein the set of predefined comparison criteria comprise at least two parameters, one parameter for comparison with an absolute difference of document rank between the requesting document and the particular document, and another parameter for comparison with a ratio of document rank between the requesting document and the particular document.

49. The computer program product of claim 40, wherein a document is a temporary redirect page comprising a document content, a source document address, and a target document address.

50. A computer program product for use in conjunction with a computer system, the computer program product comprising a computer readable storage medium and a computer program mechanism embedded therein, the computer program mechanism comprising:

instructions for constructing a plurality of tables, each table corresponding to a portion of a document address space, storing information identifying documents having a same document identifier and each identified document having an associated document rank;

instructions for receiving a newly crawled document, such document characterized by a document identifier and a document rank;

instructions for reading information stored in the plurality of tables to identify a set of documents sharing the document identifier of the newly crawled document, and ascertaining an original representative document for the identified set of documents;

instructions for updating the information stored in at least one of the tables in accordance with the document ranks of the identified set of documents and the newly crawled document;

instructions for determining a representative document for the newly crawled document and the identified set of documents;

instructions for indexing the representative document when said representative document is the newly crawled document; and

instructions for repeating the receiving, reading, updating, determining and indexing operations with respect to a plurality of newly crawled documents, each of which shares a respective document identifier with a respective set of documents, such that at least some of the newly crawled documents are determined to be representative documents and are indexed.

51. The computer program product of claim 50, wherein information identifying the identified set of documents, including a particular document serving as the original representative document of the identified set, is stored in one or more tables.

52. The computer program product of claim 51, wherein the determining includes comparing the document rank of the newly crawled document with that of the particular document from the identified set in accordance with a set of predefined comparison criteria;

selecting the newly crawled document as the representative document if the set of predefined comparison criteria are met; and

keeping the particular document as the representative document if the set of predefined comparison criteria is not met.

53. The computer program product of claim 51, wherein the set of predefined comparison criteria comprise at least two parameters, one parameter for comparison with an absolute difference of document ranks between the newly crawled document and the particular document, and another parameter for comparison with a ratio of document ranks between the newly crawled document and the particular document.

54. The computer program product of claim 50, wherein the updating includes inserting information identifying the newly crawled document into the at least one table only when a predefined insertion condition is satisfied.

55. The computer program product of claim 50, wherein the predefined insertion condition is that the document rank of the newly crawled document is higher than the document rank of at least one document in the identified set of documents.

56. A computer program product of detecting duplicate documents for use in conjunction with a computer system, the computer program product comprising a computer readable storage medium and a computer program mechanism embedded therein, the computer program mechanism comprising:

instructions for constructing a plurality of tables, each table corresponding to a segment of a document address space, storing information identifying documents having a same document identifier and each identified document having an associated document rank, wherein the plurality of tables comprise $N+1$ tables where N is an integer greater than one, wherein the $N+1$ tables comprise N tables, each generated during a respective phase of a set of N crawling phases, and a current table generated during a current one of the N crawling phases, wherein an oldest one of the N tables was generated during a previous instance of the current crawling phase;

instructions for receiving a newly crawled document, such document characterized by a document identifier and a document rank;

instructions for reading information stored in the $N+1$ tables to identify a set of documents sharing the document identifier of the newly crawled document, and ascertaining an original representative document for the identified set of documents;

instructions for updating the information stored in the current table in accordance with the document rankings of the identified set of documents and the newly crawled document;

instructions for determining a representative document for the newly crawled document and the identified set of documents;

instructions for indexing the representative document when said representative document is the newly crawled document;

instructions for repeating the receiving, reading, updating, determining and indexing operations with respect to a plurality of newly crawled documents, each of which shares a respective document identifier with a respective set of documents, such that at least some of the newly crawled documents are determined to be representative documents and are indexed; and

instructions for retiring the oldest one of the N tables upon completion of the current crawling phase.

57. The computer program product of claim 56, wherein the reading comprises reading from a merged table that stores information from a plurality of the N tables, and reading from the current table.

58. The computer program product of claim 56, wherein the identified set of documents, including a particular document serving as the original representative document of the identified set, is stored in one or more tables.

IX. Evidence Appendix

For this appeal, Appellants do not rely on any evidence submitted pursuant to §§ 1.130, 1.131, or 1.132, or other evidence entered by the Examiner.

X. Related Proceedings Appendix

Appellants are aware of no related proceedings.